Data Integration based on uniform Mapping Definitions to RDF

Anastasia Dimou, Tom De Nies, Miel Vander Sande, Ruben Verborgh, Erik Mannens, Rik Van De Walle

<firstname>.<lastname>@ugent.be
Ghent University - iMinds - Multimedia Lab

Type of presentation

Research contribution

Summary of the presentation

Handling Big Data is often hampered by integration challenges, especially when speed, efficiency and accuracy are critical. In Semantic Web, integration is achieved by aligning multiple representations of the same entities, appearing within distributed heterogeneous sources and by annotating data using same vocabularies. However, mapping different sources to RDF separately, generating multiple URIs for distinct instances of same entities, hampers advanced and automated data fusion and prevents benefiting from data semantic representation. We introduce a unified knowledge framework, created already at mapping level, to enable a successful business ecosystem built on top of enriched and combined datasets, proving that the integration of data from multiple heterogeneous sources, produces added value.

Contributors and their short CVs

Anastasia Dimou

Anastasia Dimou is a scientific researcher at Multimedia Lab, Ghent University, active in the research of (Linked) Open Data and Semantic Web technologies. She graduated from two Masters: Master in technologies for e-Government from University of Trento, Italy (UNITN) in 2009 and Master in Web Science from Aristotle University of Thessaloniki (AUTH), Greece in 2012. Her Master thesis, “Exploring Scientific Knowledge using Linked Data”, dealt with the transformation of scientific Classifications into Simple Knowledge Organization System (SKOS) and the emergence of mappings between the distinct scientific Classification Schemes. Her main interest and expertise is related to knowledge representation, integration and management. She is focused on mapping and interlinking data to semantically annotate and integrate knowledge while and after publishing Linked Data.

Tom De Nies

Tom De Nies obtained his Master's degree in Computer Science Engineering in 2010, at Ghent University, where he now works as a senior researcher at the department of Electronics and Information Systems, at iMinds - Multimedia Lab. His research is centered around automatic assessment of the value and trustworthiness of content on the Web, based on enriched metadata and provenance. In 2012 and 2013, he was a member of the W3C Provenance Working Group, where he contributed to the PROV standard. More specifically, he was one of the authors of the PROV-Constraints Recommendation and lead editor of the PROV-Dictionary Note. Git2PROV, his work in collaboration with VU University Amsterdam on exposing the provenance in version control systems received the ‘best demonstration’ award at ISWC 2013. Since 2014, he also chairs METHOD: the International Workshop on Methods for Establishing Trust of (Open) Data.

Miel Vander Sande

In 2008 Miel Vander Sande graduated as Bachelor in Multimedia and Communication technology and in 2010 as Master in Industrial Engineering: ICT. He wrote his Master thesis in collaboration with the University of Valencia, Spain about analysing and visualising RFID tracking data. After that, he concluded his education with a teachers degree in Informatics. Since september 2011, Miel joined UGent-iMinds in the research group Multimedia Lab (http://mmlab.be) as a researcher. His main interest and expertise are (linked open) data publishing (a.o. in the context of Open Knowledge Foundation), versioning, and querying on a Read/Write Web. He is active as Open Data activist in the Belgian and European community, supporting policy making and the organisation of events (such as App contests). Furthermore, he participates in the Semantic Web research community by organising several international workshops (WaSABi, SemDev, and NoISE) and as member of the W3C Linked Data Platform Working Group. Currently, Miel is active in multiple Flemish government projects for creating Linked Open Data ecosystems. Additionally, he participates in a Flemish innovation project for the digitalisation of book publishers and eBooks.

Ruben Verborgh

Ruben Verborgh is a researcher in semantic hypermedia at Ghent University – iMinds, Belgium, where he obtained his PhD in Computer Science in 2014. He explores the connection between Semantic Web technologies and the Web's architectural properties, with the ultimate goal of building more intelligent clients. Along the way, he became fascinated by Linked Data, REST/hypermedia, Web APIs, and related technologies. He's a co-author of two books on Linked Data, and has written more than 100 publications on Web-related topics for international conferences and journals.

Erik Mannens

Erik Mannens is Professor at MMLab's KNoWS group and experienced Research and Project Manager at iMinds Media Technologies Dept where he has successfully managed +30 projects. He received his PhD degree in Computer Science Engineering (2011) at UGent and his Master’s degree in Computer Science (1995) at K.U. Leuven University. Before joining MMLab, he was a software engineering consultant and Java architect for over a decade. His major expertise is centered around metadata modeling, semantic web technologies, broadcasting workflows, iDTV and web development in general. He was co-chair of the W3C Media Fragments Working Group and actively participating in other W3C’s semantic web standardization activities (Media Annotations, Provenance, Hydra, Linked Data Platform, and eGovernment). Since 2008 Erik is paving the Open Data path in Flanders. He stood at the cradle of the first Hackatons and is a founding member of the Open Knowledge Foundation. (Belgian Chapter). Since then, he is frequently invited as Open Data evangelist at national and international events. Currently he actively participates in W3C’s eGov and Data On The Web working groups. On all of these subjects he has published +100 papers and book chapters. He is also member of the technical committee and/or organizing committee of several high level journals and conferences.

Data Integration based on uniform Mapping Definitions to RDF

Anastasia Dimou, Tom De Nies, Miel Vander Sande, Ruben Verborgh, Erik Mannens, Rik Van De Walle
<firstname>.<lastname>@ugent.be
Ghent University - iMinds - Multimedia Lab

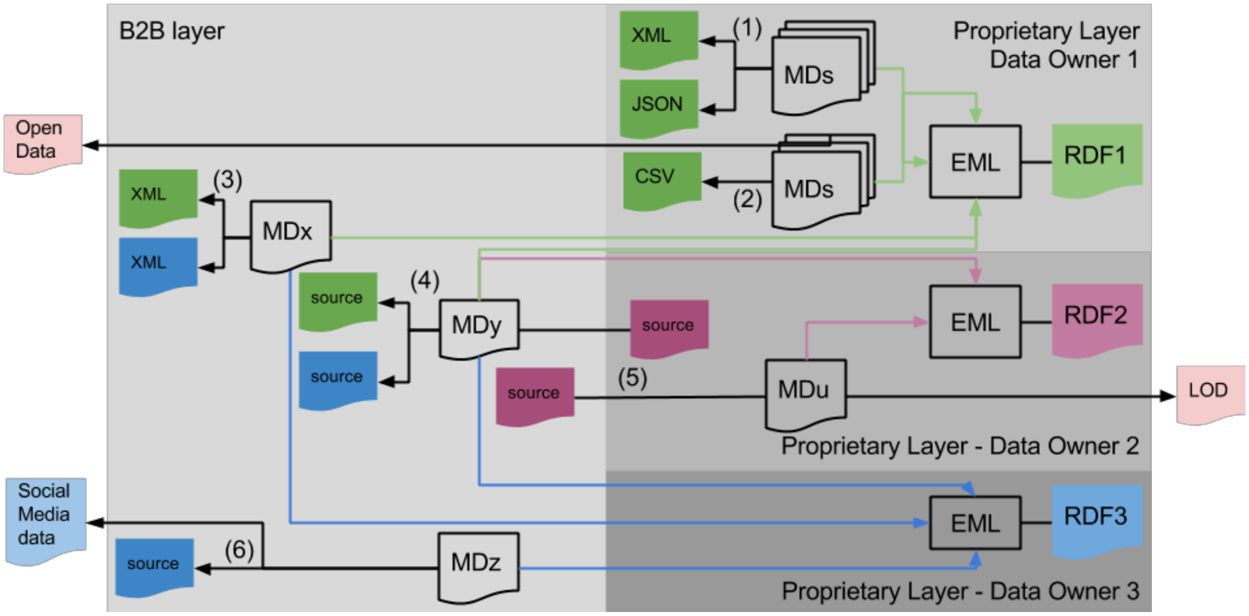
As the Web evolves in an integrated and interlinked knowledge space thanks to the growing amount of published data, companies have been overwhelmed with this continuously increasing amount of data. Big Data open major opportunities for offering Data-as-a-Service (DaaS) to third party players and thus new data-driven business applications. Creating value from data requires efficiently handling huge volumes of dispersed information which is often hampered by challenges related to data integration and management. Data owner have different kinds of data, thus different levels of knowledge, different purpose of data usage and are eager to incorporate complementary data into their core data management process. Nevertheless, the integration of new data sources, in practice, is very cumbersome, requires a substantial manual effort and often leads to rather long IT development cycles. This occurs mostly because the data is not originally in common formats or structures nor share same schema or are annotated with common vocabularies, thus additional algorithms and business rules are required to validate suspect matches or aggregate conflicting data.

The problem of identifying and aligning multiple but possibly (slightly) different representations of the same real-world entities, appearing within multiple distributed heterogeneous sources, aggravates when those sources are of big size, and especially when speed, efficiency and accuracy is very critical. In Semantic Web, those entities are ideally identified by the same URI (acting as the entity's unique identifier). However, due to the high degree of mismatches between the structure and content of data across different sources and the variety in vocabularies, aligning the same entities to automatically integrating these data sources is a challenging and meticulous procedure. Therefore, the need for solutions to discover, retrieve and effectively integrate data emerges. Existing technologies would map those data sources to the RDF data model separately, generating multiple URIs for each distinct instance of the entity. Such an approach conflicts with the Linked Data principles, hampers advanced and automated data fusion and in the end, prevents users from actually and directly benefiting from the semantic representation of the data.

Our approach focuses on creating a combined, unified knowledge framework already at mapping level, ensuring data veracity and conformity that contributes to efficient data integration and management. Our use cases focus on aggregation and disaggregation of high-variety, multi-source and heterogeneous data to interchange and integrate SME's data. The mismatches and conflicts of entities alignments are validated by determining the data provenance and by amending data taking into consideration crowdsourcing results that form a dataset on their own. We materialize our approach at the COMBUST platform[1], considering RDF Mapping Language (RML)[2] to define the Mapping Definitions [1]. With our approach, we aim to enable a successful business ecosystem built on top of enriched and combined datasets, proving that the integration of data from multiple heterogeneous sources, actually produces added value.

[1] http://www.iminds.be/en/projects/2015/03/11/combust
[2] http://rml.io

Data integration at mapping level is achieved thanks to the mapping definitions applied to the original data. Our approach allows to reuse the same mapping language to incrementally (i) ***integrate*** data owner's heterogeneous resources; (ii) ***enrich*** them with information derived from different sources available on the Web, e.g. Open Data or Social Media; (iii) ***incorporate*** information from other data owners data shared over the B2B layer, leading to unprompted data ***interchange***; and (iv) resolve mismatches, ambiguous alignments and conflicting information derived from different data owners taking advantage of the data provenance or the crowdsourcing results. This is achieved because of (i) the ***uniform and interoperable mapping definitions***; and (ii) the ***robust cross-references to the mapping definitions and interlinking***.



For all heterogeneous sources data owners can define ***uniform*** mappings to RDF independently of the source's type or format, Fig. (1), which are ***reused*** across data sharing the same structure, Fig. (3). Those mapping definitions are ***interoperable*** across different data owners Extract-Map-Load (EML) systems. This way *same URIs are assigned for equivalent entities*, ***avoiding duplicates*** generation and data is annotated with the *same schema* (vocabularies and ontologies). Mapping Definitions is ***shared*** and used to annotate proprietary, or not, data which get ***enriched*** with information derived from complementary shared sources, Fig. (5). Moreover, uniform mapping definitions allow references to other openly available data, e.g. Open Data sources, Fig. (2), or even data derived from social media, Fig. (6), e.g.twitter. This way, not only information is ***integrated***, but also relationships among data originally residing in different sources are attained (***interlinked***). The pattern that generates a resource is uniquely defined and this mapping definition is ***referred*** every other time this resource is required, Fig. (4). Thus, modifications to the patterns, or data values appearing in the patterns that generate the URIs, are ***propagated*** to every other reference of the resource, keeping the different views ***synchronized***. At the end, each data owner executes only the relevant Mapping Definitions, periodically or when updated, to acquire the desired view of the data.

[1] Anastasia Dimou, Miel Vander Sande, Pieter Colpaert, Ruben Verborgh, Erik Mannens, Rik Van De Walle. RML: A Generic Language for Integrated RDF Mappings of Heterogeneous Data. Proceedings of the 7th Workshop on Linked Data on the Web, WWW14